DOCUMENT RESUME

ED 094 337 CS 001 228

AUTHOR Walter, Richard F.
TITLE Historical Overview of the Cloze Procedure.
PUB DATE 74
NOTE 33p.; Study conducted at Kean College of New Jersey

EDRS PRICE MF-$0.75 HC-$1.85 PLUS POSTAGE
DESCRIPTORS *Cloze Procedure; *Historical Reviews; Literature Reviews; *Readability; Reading; *Reading Comprehension; Reading Level; *Reading Research; Reading Skills

ABSTRACT

The use of the cloze procedure as a method of determining relative readability was first introduced by Wilson Taylor, who contended that the cloze procedure was a superior method of determining readability since it measured a facet of readability called concept load. Authorities have accepted cloze tests as valid and reliable instruments capable of measuring general comprehension skills. Agreement now exists as to the accepted methodology of cloze test construction. A deletion system of every fifth word with standardized blank length has been accepted, although recent studies have submitted that blank length is not crucial to the construction of cloze tests. Three studies indicated the suitability of the cloze procedure as a method of determining the instructional levels of elementary school pupils. In reviewing the literature, many research weaknesses become apparent. Many researchers failed to mention the procedures which were employed in cloze test construction. Deletion system and scoring procedures are often described inadequately. More needs to be known about the construct validity of cloze tests. Several affective components operating during the completion of a cloze test also need to be investigated. (WR)

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

PERMISSION TO REPRODUCE THIS COPYRIGHTED MATERIAL HAS BEEN GRANTED BY
Richard B. Walter

TO ERIC AND ORGANIZATIONS OPERATING UNDER AGREEMENTS WITH THE NATIONAL INSTITUTE OF EDUCATION. FURTHER REPRODUCTION OUTSIDE THE ERIC SYSTEM REQUIRES PERMISSION OF THE COPYRIGHT OWNER.

ED 094337

Historical Overview of the Cloze Procedure

by Richard B. Walter
Kean College of New Jersey

Many of the innovations in reading are not based upon entirely new concepts. Rather, they are relatively similar to methods which have preceded them. An example is the collection of processes used to determine readability. All are methodologically applied to the reading material without a reader being involved. Most of the methods of determining readability are concerned with the presence or absence of specific words in the reading selection, and with average sentence length. Thus, it was with interest that the reading world viewed the newly introduced cloze procedure. Here was a method that determined the difficulty of reading material by ranking the selections in order of difficulty. It also depended upon the presence of a reader.

The cloze procedure was developed by Wilson L. Taylor in 1953 when he was doing graduate work at the University of Illinois. Taylor applied the statistical principles of randomization to the mutilation of a prose passage. This mathematical deletion of words (that is, every n^{th} word) was used as an exercise to determine the success a reader would have in supplying the missing word. Taylor designated this the "cloze procedure."[1]

Taylor stated:

> It is pronounced like the word "close" and is derived from "closure." The last term is one Gestalt psychology applies

---

[1]Wilson L. Taylor, "Cloze Procedure: A New Tool for Measuring Readability," Journalism Quarterly, XXX (Fall, 1953), 414.

CS 001 228

> to the human tendency to complete a familiar but not-quite-finished pattern -- to "see" a broken circle as a whole one, for example, by mentally closing up the gaps.[2]

Twelve years later, Weaver took issue with Taylor's analogy and claimed that the designation "cloze" was misleading since many of the common constructs of Gestalt psychology cannot be easily applied to verbal situations.[3] Nonetheless, the term survives.

Models have long been popular in education. The profession has used physiological and psychological models; it is presently being influenced by a computer and space age model, that of systems analysis. Taylor elected to use an electronic model for his illustrative definition of the cloze procedure. He stated:

> Cloze procedure may be defined as: a method of interrupting a message from a "transmitter" (Writer or speaker), mutilating its language patterns by deleting parts, and so administering it to "receivers" (readers or listeners) so that their attempt to make the patterns whole again potentially yield a considerable number of cloze units.[4]

Taylor defined cloze units as:

> Any single occurrence of a successful attempt to reproduce accurately a part deleted from a "message" (any language product) by deciding, from the context that remains, what the missing part should be.[5]

---

[2] Ibid., p. 415.

[3] Wendell W. Weaver, "Theoretical Aspects of Cloze Procedure," Fourteenth Yearbook of the National Reading Conference, (Milwaukee: National Reading Conference, Inc., 1965), pp. 115-32.

[4] Taylor, op. cit., p. 416.

[5] Ibid.

Initially, Taylor applied the cloze procedure to the task of determining the relative difficulty of a selection of reading material. As previously stated, this was a rather unique approach to readability, for, as Anderson stated, "Cloze as a readability technique involves the reader."[6]

The basic theory behind the use of the cloze procedure as a method of determining readability was simply stated by Hafner: "The individual's choice [of words] is an index of his ability to comprehend reading matter."[7]

A conclusion of Taylor's research was that the cloze procedure was a more valid predictor of reading difficulty than either the Dale-Chall or the Flesch formula. Taylor found that these formulae could be "fooled" by authors' style and selected specific examples from literature to demonstrate this. Passages were selected from the works of Gertrude Stein and James Joyce. The writings of Stein and Joyce have, as a common feature, passages which are easily read in terms of the level of vocabulary, but with a high concept load. Taylor concluded that the Dale-Chall and the Flesch formulae were not sensitive enough to many aspects of reading, since the application of these two formulae yielded very low levels of

---

[6]J. Anderson, "Research in Readability for the Classroom Teacher," Journal of Reading, X (February, 1967), 291-99.

[7]L. E. Hafner, "Cloze Procedure," Journal of Reading, IX (May, 1966), 415-21.

readability when applied to the passages written by Stein and Joyce. Taylor felt that the cloze procedure did not suffer from this deficit and stated:

> Potentially important, it seems, is the fact that a cloze score seems to be a measure of the aggregate influence of all factors which interact to affect the degree of correspondence between the language patterns of transmitter and receiver.[8]

Having confirmed his belief that the cloze procedure is a valid indicator of readability, Taylor submitted that the cloze procedure should not be limited to this particular aspect of reading. He wrote:

> ... as such, its potential usefulness is by no means confined either to readability or to the reading abilities of individuals.[9]

Several years later, Taylor again wrote of the cloze procedure, this time expanding upon his previous premonition of its potential usefulness. Referring to his initial article, he stated:

> At that time, this writer tended to regard cloze procedure mainly as a "new and better" way of determining the readability of printed English prose. Even then, however, there were indications that the technique was not limited to readability matters. And now--two years and many experiments later--the cloze method appears to possess a very large number of potential research uses, both practical and theoretical.[10]

Taylor further stated:

> The technique seems to be an effective guage of "individual differences" in the comprehension of readers ....[11]

By this time, the cloze procedure had attracted the attention of other investigators. Weintraub reported on the surveys of the literature which had been done by Rankin and observed that while the initial review of the literature done by Rankin in 1959 contains a bibliography of 12 studies, a

---

[8] Taylor, op. cit., p. 432

[9] Ibid.

[10] Wilson L. Taylor, "Recent Development in the Use of the Cloze Procedure." Journalism Quarterly, XXXIII (Winter, 1956), 42

[11] Ibid.

similar review published in 1965 contains a bibliography of almost 50 articles.[12]

Bormuth, a most prolific writer on the topic of cloze procedure, discussed the cloze procedure as a method of readability in 1963.[13]

Hafner followed with an article on the implications of cloze in 1965.[14] Gallant obtained correlations between cloze scores and grade equivalent scores on comparable forms of a standardized reading test, one rewritten as a cloze test.[15] Gallant's conclusions are to be interpreted cautiously, as she assumed the cloze procedure to be too complex a procedure for first graders and modified it into a multiple-choice task. Thus, by offering possible responses, Gallant deviated from one of the basic ground rules of the cloze procedure--that the subject supplies the response of his choice. Gallant supplied no rationale for the method of selecting the non-correct responses which appeared on the test. Since, in almost all cases, one of the responses would not fit the blank due to a difference in language patterns, the results would seem to have little application to the cloze procedure.

Greene reported upon the use of the cloze procedure in the primary grades

---

[12]Samuel Weintraub, "Cloze Procedure," Reading Teacher, XXI (March, 1968,)567.

[13]John Bormuth, "Cloze as a Measure of Readability," Proceedings of the International Reading Association, VIII (Newark, Delaware: International Reading Association, 1963), 131-34.

[14]L.E. Hafner, "Implications of Cloze." Fourteenth Yearbook of the National Reading Conference, Inc., 1965), pp. 134-45.

[15]Ruth Gallant, "Use of Cloze as a Measure of Readability in the Primary Grades," Proceedings of the International Reading Association, X (Newark, Delaware: International Reading Association, 1965), 286-87.

in 1965[16] and discussed many potential uses of the cloze procedure in 1967.[17]

It was during this period of time that many researchers began studying specific aspects of the cloze procedure. The more significant studies will be discussed beginning with cloze test construction.

Cloze Test Construction

Taylor's initial article on the cloze procedure contained a suggested methodology for the construction of a cloze test. This test was to be used to determine the readability, or relative difficulty, of a specific selection of reading material. Taylor's directions were as follows:

1. Depleting an equal number of words from each passage by some essentially random counting-out system. Such a system is based on a table of random numbers or else is simply counted out every n^{th} word (every fifth one, for example) without any regard for the functions or meanings of specific words.

2. Reproducing each mutilated passage with a blank of some standard length (so the length will not influence guessing) in place of every missing word.

3. Giving copies of all reproduced passages to all subjects—or to equal numbers of randomly selected subjects—in a sample group representative of the population in question.

4. Asking all subjects to try to fill in all blanks by guessing, from the context of remaining words, what the missing words should be.

5. Totaling for each passage separately the number of times original words were correctly replaced, and considering these totals as readability scores.

---

[16] Frank Greene, "Modification of the Cloze Procedure and Changes in Reading Test Performances," Journal of Educational Measurement, II (1965), 213-27.

[17] Frank Greene, "Cloze Symposium," Educational Resources Information Center Document ED 016 581 (Bethesda, Maryland: ERIC Document Reproduction Service, 1967), Abstract.

6. Contrasting the cloze totals of various passages: the passage with the highest score was considered "most readable", the one with the second highest score, next most readable, etc., pending the outcome of statistical tests of significance of the differences observed.[18]

Deletion Systems

One of the first aspects of cloze test construction to come under investigation was the deletion rate of words. In his initial article, as cited above, Taylor suggested a deletion rate of every fifth word. There was little evidence at that time to support Taylor's decision, which, it seems, was completely arbitrary.

Taylor provided the rationale in his second article on the cloze procedure. His statement was as follows: "Also, it appears that an every-fifth-word deletion system spaces blanks as far apart as they need to be."[19]

The rationale for this decision was included in a footnote in which Taylor indicated that Madow, a statistician, had "...verified that a subject's performance on successive blanks created by an every-fifth-word deletion is statistically independent."[20] The problem of statistical dependency occurs in deletion systems of less than every-fifth-word. The rationale of using a deletion system of less than every fifth word is based upon the benefits gained by obtaining a greater number of cloze units from a given paragraph. This reduces the length of the test, length being a point of concern necessary to insure the reduction of sampling error and to insure reliability.

---

[18]Taylor, "Cloze Procedure," op.cit., p.416.

[19]Taylor, "Recent Developments," op. cit., p.48

[20]Ibid.

MacGinitie designed a study to investigate the effects of a less than every-fifth-word deletion system. He concluded that:

> ...no statistically significant difference was found in the difficulty of restoring omitted words when every 24th, 12th, or 6th word was omitted; but omitting every 3rd word made restoration more difficult.[21]

MacGinitie's concern with less than every-fifth-word deletion systems was not matched by most investigators. Many research studies have investigated systems which allow for more than four words between blanks.

Culhane based his requirements for a deletion system on the types of materials being read. He recommended the every-fifth-word deletion system for narrative materials but felt that an every-tenth-word count was appropriate for textual materials that were fact-laden.[22] Potter concurred with this reasoning and even suggested a deletion system of one word in every twelve for certain purposes.[23]

McLeod and Anderson attempted to use a deletion rate of every-eighth-word and found that certain blanks evoked ambiguous responses. These blanks were replaced, resulting in a loss of randomization, and this procedure was dismissed as being unsuccessful.[24]

---

[21]W.H.MacGinitie, "Contextual Constraint in English Prose Paragraphs," Journal of Psychology. LI (1961), 125.

[22]J.F. Culhane, "Cloze Procedure and Comprehension," Reading Teacher, XXIII (February, 1970), 410-13.

[23]Thomas C. Potter, "A Taxonomy of Cloze Research, Part I: Readability and Reading Comprehension," Educational Resources Information Center Document ED 022 644 (Bethesda, Maryland: ERIC Document Reproduction Service, 1968), Abstract.

[24]J. McLeod and J. Anderson, "Readability Assessment and Word Redundancy of Printed Speech," Psychological Reports, XXVIII (February, 1966), 35-38.

In view of the findings of the previously mentioned investigations, it is particularly interesting to note a study by MacGinitie which involved 600 college students. MacGinitie used deletion systems of one in six, one in twelve, and one in twenty-four. He found no significant differences among the cloze test scores.[25]

A question also arose as to the desirability of making lexical rather than structural deletions. It is conceivable that cloze tests can be constructed by deleting a specific class of words such as nouns or adjectives. This technique seems justified in the event the investigator wishes to assess the language competencies of a group of subjects. It would seem that such a technique would be unsuitable for purposes of assessing readability or comprehension skills. The deletion of a specific class of words and the frequency with which they occur in a paragraph, in and of itself, might be an influencing factor which could affect a subject's cloze score by altering the difficulty of the passage. Taylor researched this deletion technique and concluded that: "Findings up to now indicate that...there seems to be little advantage in preclassifying words and limiting deletions to them...."[26]

Further, it would seem that once the randomized deletion rate is separated from the cloze procedure, the loss of objectivity would reduce the task to little more than sentence completion. Also, one of the most important qualities of the cloze procedure, the fact that the selection process of test items is completely void of subjectivity would be lost.

Potter's statement, as a result of an extensive review of the cloze procedure best summarizes present thought on this matter. He wrote: "The practice of selecting

---

[25]MacGinitie, op. cit., pp. 121-30.

[26]Taylor, loc. cit.

cloze deletions on other than a pre-determined, mechanical formula seems questionable since a random sample of the test involved is not obtained."[27]

Having established the most successful sustem to be used, attention will now be given to the total number of deletions necessary for inclusion in a cloze test.

In his second article, Taylor suggested a passage length of 250 words.[28] Potter proposed a series of guidelines for cloze test construction and stated that:

> ...(2) not more than 20 words in every hundred are deleted, (3) paragraph length be at least 250 words, (4) at least 50 words are deleted in order to assure adequate sampling of passages.[29]

Rankin, in a study of students in grades four to eight, also used a 250-word passage with every fifth word deleted.[30] There is little debate in the literature as to the appropriateness of the 250-word length. Most investigators have accepted this standard.

Blank Length

While the number of blanks does not seem to be an issue, the length of the blanks has been given some thought. Early investigators have generally accepted the logic of Taylor's statement that blanks should be of a predetermined length (usually ten spaces) so as to curtail the subject's attempts to guess the correct

---

[27]Potter, loc. cit.

[28]Taylor, "Recent Developments," op. cit., p. 48.

[29]Potter, loc. cit.

[30]Earl F. Rankin, "Grade Level Interpretation of Cloze Readability Scores," Educational Resources Information Center Document ED 046 057 (Bethesda, Maryland: ERIC Document Reproduction Service, 1970), Abstract.

word by seeing if it "fits" the length of the blank.[31] Culhane agreed with this, but specified that a consistent blank length was important only if the exercise was to be used as a test. Culhane felt that unintentional clues of this type would affect pupil performance in a testing situation.[32]

Anderson designed a study to investigate the effects of blank length on performance and concluded that:

> No significant difference was found between the mean scores of subjects on cloze tests using blanks of a uniform length and the mean scores of subjects on cloze tests using blanks of the same length as the deleted words. Both versions of the cloze tests were equally valid as measures of general reading comprehension.[33]

The question of blank length is of practical as well as academic interest. If further research will support Anderson's findings, construction of cloze tests will be greatly simplified. In order to construct a cloze test utilizing a uniform blank length procedure, it is necessary to reprint the appropriate passage, rearranging the sentence spacing so as to accomodate the blank length of ten spaces. This necessitates either printing costs or typing effort. Even when a primary typewriter is used, the visual advantages of print size common to primary reading materials is lost.

The question of blank length has not received as much attention as the responses which fill the blanks. Much of the early research on the cloze procedure was addressed to the problem of what would constitute an appropriate cloze response.

---

[31]Taylor, "Cloze Procedures," op. cit., p. 416.

[32]Culhane, op. cit., pp. 410-413.

[33]Jonathan Anderson, "A Report of Research on Comprehension in Reading," Educational Resources Information Center Document ED 045 323 (Bethesda, Maryland: ERIC Document Reproduction Service, 1970), Abstract.

Acceptable Cloze Test Responses

Taylor was explicit in his directions pertaining to cloze responses. He recommended "totaling for each passage separately the number of times original words were correctly replaced and considering these totals as readability scores."[34]

In his second article, published in 1956, Taylor again addressed himself to the problem of correct responses:

> Each time the subject correctly guesses a missing word, he scores one point; his "cloze score" for any particular passage is simply the total number of missing words that he guesses correctly.[35]

Thus, each cloze passage yields a raw score based upon the principle of one point per correct response.

The problem of equating scores on cloze passages arises when passages differ in length. To resolve this problem, raw scores are usually converted to percentage scores.

Taylor's stipulation to accept only exact word replacement responses removes the element of subjective judgement from scoring the tests. As desirable a feature as this may be, many researchers began investigating the effects that accepting alternate responses had upon the validity of a cloze test.

One of the earliest of the alternate responses to be investigated was synonyms. The agreement to accept synonyms is logical. Our language lacks specificity: we use many words to describe a particular object. Taylor felt that scoring synonyms did not justify the extra effort involved. Taylor contended that there was

---

[34]Taylor, "Cloze Procedures," op.cit., p. 416.

[35]Taylor, "Recent Developments," op. cit., p. 43.

> ... no advantage to putting oneself to the trouble of judging and scoring synonyms .... Findings up to now indicate that the easiest ways of applying cloze procedure may be best for most uses.[36]

Culhane supported Taylor's beliefs and wrote:

> It is possible that synonyms may be scored as correct, but this would make scoring cumbersome and lead to arbitrary decisions regarding the worth of the synonyms as a replacement.[37]

In a study which compared the exact word replacement method with the synonym replacement method, Ruddell found no significant differences between the two scoring methods in terms of the validity and reliability of the tests.[38] Gallant decided, as a result of her study, that accepting substitute words which approximated the meaning of the word deleted and agreed in both person and tense was not statistically significant even though higher correlations were obtained using the substitute scoring method.[39]

Since not all elementary school children are proficient spellers, the problem of how to interpret misspelled words was acknowledged by Bormuth. He suggested accepting misspelled words when the response is otherwise correct, and when the misspelling does not result in the spelling of another word which does not fit the syntactical context of the cloze blank.[40] Bormuth also conducted

---

[36] Ibid., p. 48.

[37] Culhane, op. cit., p. 412.

[38] Robert Ruddell, "A Study of the Cloze Comprehension Technique in Relation to Structurally Controlled Reading Materials," Proceedings of the International Reading Association, IX (Newark, Delaware: International Reading Association, 1964), 298-302.

[39] Ruth Gallant, "An Investigation of the Use of Cloze Tests as a Measure of Readability of Materials for the Primary Grades," (unpublished doctoral dissertation, Indiana University, 1964).

[40] John Bormuth, "The Implications and Use of the Cloze Procedure in the Evaluation of Instructional Programs," Educational Resources Information Center Document ED 012 074 (Bethesda, Maryland: ERIC Document Reproduction Service, 1967), Abstract.

various studies to determine the effects of alternate responses upon validity and concluded that the most economical and objective methods of scoring are to be preferred. The exact word replacement method, he decided, yields the most valid results.[41] Jongsma addressed this problem after an extensive survey of the literature on the cloze procedure and observed:

> Thus it appears that the literature consistently shows the scoring of exact word replacements to be the most objective, efficient, and useful scoring system to use with the cloze procedure.[42]

The preceding sections contain a discussion of some of the techniques and problems of cloze test construction. The following section examines the question of whether or not the cloze procedure and cloze tests are valid and reliable measures of readability.

Validity and Reliability of Cloze Tests as Indicators of Readability

In his initial publication on cloze procedure, Taylor supported the validity of the cloze test by comparing it with the Dale-Chall and the Flesch readability formulae. In this study, Taylor demonstrated that standard readability formulae can be "fooled" by applying these formulae to passages written by James Joyce and Gertrude Stein. These authors' writings are ranked as more difficult by the cloze procedure than by the Dale-Chall and the Flesch formulae. Taylor explained this by pointing out that the Dale-Chall and the Flesch formulae cannot evaluate

---

[41] Ibid., p. 20.

[42] Eugene A. Jongsma, "The Cloze Procedure: A Review of the Research," Occasional Papers in Reading (Bloomington, Indiana: Indiana University School of Education, n.d.), p. 7.

comprehension difficulty as can the cloze procedure which, Taylor felt, measures the factor of concept difficulty. The cloze procedure, he contended, does not suffer from the inadequacies of a pre-selected word list.[43]

A further point to be made in support of the cloze procedure can be found in the method used to determine the readability of a selection of reading material. All too often, the readability level of a book is found by taking three sample selections and computing the arithmetic mean of these samples. Inaccuracies occur when these samples are averaged. The mean readability score is often far below the score of the most difficult selection. For example, the mean score of the following readability levels 2.0, 5.0, and 7.0 is 4.7. Some might conclude that this particular selection is suitable for fourth grade pupils who are at grade level at the end of the year and that it is certainly suitable for most fifth graders. However, in order to be capable of reading the entire selection, a fifth grade pupil is going to have to cope with some seventh grade reading material.

The cloze procedure, which does not make use of arithmetical averages, seems to yield a more valid indication of readability. Further, most readability formulae are passive: they are formulae applied to the selection without input from the pupil. The cloze procedure might be termed a more active method since it involves the reader as he is reading the material.

Froese compared the validity of the Dale-Chall readability formula for sixth grade science textbooks with a cloze test over the same material. His major conclusions were:

---

[43]Taylor, "Cloze Procedure," op. cit., pp. 414-38.

> (1) Cloze tests were reliable measures of language difficulty and (2) the Dale-Chall readability formula is not a valid measure of sixth grade science textbook materials when the cloze procedure is used as a criterion.[44]

In a study to determine grade level readability scores, Fry compared the Fry Readability Graph with the Spache formula, the cloze procedure, and oral reading using seven primary level textbooks. He reported high correlations for all four methods and concluded that "the cloze method was the most accurate and the most capable of making fine distinctions."[45] However, Fry felt that the cloze procedure had limited use, since it required the presence of a group of subjects to read the selections.

Weintraub also investigated the question of cloze validity and reliability. His conclusions, based on a review of several studies, were:

> . . .(1) that the cloze procedure provided a more reliable measure of readability for non-standard reading materials than commonly used readability formulas, (2) that cloze tests are valid and reliable predictors of the difficulty level of reading materials, and (3) that scores on comprehension tests correlate highly with cloze readability scores. These findings indicated that the cloze procedure could be used effectively as a measure of readability and of comprehension.[46]

It appears that the validity of the cloze procedure is an accepted fact insofar as the measurement of readability is concerned. The following discussion

---

[44]Victor Froese, "Cloze Readability Versus the Dale-Chall Formula," Educational Resources Information Center Document ED 051 975 (Bethesda, Maryland: ERIC Document Reproduction Service, 1971), Abstract.

[45]Edward B. Fry, "The Readability Graph Validated at Primary Levels," Educational Resources Information Center Document ED 232 565 (Bethesda, Maryland: ERIC Document Reproduction Service, 1962), Abstract.

[46]Samuel Weintraub, "The Cloze Procedure," Educational Resources Information Center Document ED 027 145, (Bethesda, Maryland: ERIC Document Reproducation Service, 1968), Abstract.

centers around the use of the cloze procedure as a valid and reliable indicator of comprehension skills.

Validity and Reliability of Cloze Tests as Indicators of General Comprehension

Cloze tests appear to be a valid measuring device based upon the following logic relating to face validity.

Face Validity of the Cloze Procedure

Both standard achievement and teacher-made tests in reading most often rely upon the use of multiple-choice or sentence completion items to test reading comprehension skills. A multiple-choice test readily lends itself to the process of elimination in selecting correct responses. Many of the responses to the stem of the question are easily eliminated by the subject as he considers the subtle clues inherent in the construction of multiple-choice questions. Differences in verb agreement, redundancy of sentence style, syntactical mis-matching and illogical alternatives furnish the more sophisticated test-taker with a variety of clues which guide his choice of responses. Subjects often disregard what is obviously wrong in an effort to determine what is probably right. Further, the statistical benefits of guessing should not be overlooked. Bormuth offered two sets of scores in the results of one of his studies. The first set was raw test scores; the second set had been statistically corrected for guessing. Bormuth stated:

> A set of corrected multiple-choice scores was also calculated. This calculation was based upon the assumption that a subject's score is made up of two components. The number of items on which he knew the right answer, and the number of items on which he guessed correctly. It was also assumed, because there were four alternatives for each item, subjects guessed correctly by

one-fourth of the items for which they did not know the correct answers.[47]

Sentence completion test items do not suffer from the previously mentioned drawbacks. However, in both the multiple-choice and sentence completion type of test question, vocabulary control and concept control of the questions are factors to be considered. Generally, test items are derived from the text but often contain grammatical constructions and vocabulary which are unfamiliar to the subject taking the test. The cloze procedure, which draws from the actual material itself for test items, inserts no added vocabulary or changes in grammatical constructions. Rather, the text is the test. The problem of a pupil comprehending the selection, but not the question, can never be encountered in the cloze procedure.

Concurrent Validity of the Cloze Procedure

Most of the attempts to prove concurrent validity have dealt with the nature of the relationships among the cloze procedure and other standardized instruments and procedures. Taylor found correlations of .76 between scores on a cloze test and a comprehension test made from the same material.[48] Bormuth reported correlations of .73 to .84 between cloze test scores and conventional test scores made over the same material and stated the correlations approached 1.00 when corrected for the unreliability of the test.[49] In a similar study, Bormuth found correlations of .92 between cloze

---

[47]John Bormuth, "Comparable Cloze and Multiple-Choice Comprehension Test Scores," Journal of Reading, X (February, 1967), 298.

[48]Taylor, "Recent Developments," op. cit., p. 45.

[49]John Bormuth, "Cloze Readability Procedure," Elementary English, XLV (April, 1968) 431.

readability scores over a passage and the difficulty of the same passage as measured by multiple-choice tests.[50] In a later study, Bormuth attempted to determine the relationship between the cloze procedure and word recognition and comprehension as separate concepts. He used all four forms of the Gray Oral Reading Paragraphs Test and found correlations of .90 to .95 between the cloze test scores and word recognition difficulties. Correlations of .91 to .96 were found between the cloze test scores and comprehension difficulties relating to the paragraphs.[51] An interesting light in which to view these findings is this: if correlations exist among cloze test scores and comprehension and word recognition difficulties, an assumption may be made that a correlation should exist between scores on a cloze test and scores on an IRI, an instrument which measures both word recognition and comprehension abilities. Literature pertaining to this point is presented in a subsequent portion of this chapter. The following section is concerned with the construct validity of the cloze procedure.

Construct Validity of the Cloze Procedure

The literature reveals little agreement on this subject. This lack of agreement seems to be due to the absence of a theoretical framework, since the processes a person goes through in completing a cloze test are not presently known. An assumption is made, and perhaps rightly so, that since the processes necessary for performance on a cloze test are not unlike the processes of comprehension itself (as evidenced by high correlation coefficients), a similarity exists. Jenkinson's study investigated this problem. She attempted to examine the process of comprehension by asking her subjects to "think aloud" while completing a cloze test. Thus, she developed a

---

[50] Ibid.

[51] Ibid.

clarification system for analyzing the process of comprehension.[52]

Kingston and Weaver applied the principles of factor analysis to the cloze procedure in an attempt to determine the proportions of variance that could be assigned to those factors which are basic to tests of reading ability, namely, vocabulary and language aptitude. A series of standardized tests, four reading cloze tests, and four listening cloze tests were administered to juniors in college. The factors which were isolated were a verbal comprehension factor; a cloze factor; and a rote memory, flexible retrieval factor.[53] The data showed that cloze tests were not related to the above three factors. Weaver and Kingston concluded that "cloze tests are related only moderately to the verbal comprehension factor."[54] Bormuth reported several criticisms of the above-mentioned study. He designed his research in an attempt to eliminate many of the limitations he found in the study done by Weaver and Kingston.

It would be appropriate to discuss the two studies by contrasting them, since they are similar in scope, and Bormuth's study is an attempt to rectify certain "wrongs" which he found in the Weaver-Kingston study. Bormuth elected to use an every-fifth-word deletion system as opposed to the selective deletion system used in the Kingston-Weaver study. Bormuth randomly selected fourth, fifth, and sixth graders, while Weaver and Kingston used college juniors. In the area of criterion measures, Bormuth constructed his own multiple-choice comprehension questions and based the tests' validity on pilot testing and professional judgement; Weaver and Kingston used standardized tests as their

---

[52]Marion Jenkinson, "Selected Processes and Difficulties of Reading Comprehension" (unpublished doctoral dissertation, University of Chicago, 1957).

[53]W.W. Weaver and A.J. Kingston, "A Factor Analysis of the Cloze Procedure and Other Measures of Reading and Language Ability," Journal of Communication, XIII (December, 1962), 252-61.

[54]Ibid., p. 259.

criterion measures. Bormuth concluded that "... one factor accounted for the preponderance of the variance...."[55] Bormuth found "... little diffuculty applying the name 'reading comprehension' to that factor."[56]

The major studies which dealt with construct validity have been presented. To this date, there is no research to prove positively the construct validity of the cloze procedure. It is appropriate to assume the existence of a relationship based upon the high, positive correlations that have been found between cloze tests and many reading comprehension tests.

One of the first studies to investigate the validity and reliability of cloze tests as a measuring tool for general comprehension was Jenkinson's study of 1957. Jenkinson made deletions on the basis of word difficulty. Deletion rates varied from every third word to every tenth word. Correlations between the cloze test and the Cooperative Reading Test were .78 with the vocabulary section and .73 with the comprehension section.[57]

Hafner found correlations between cloze test scores and performance on the Michigan Vocabulary Test (for college students). A correlation coefficient of .56 was found.[58]

Ruddell's study used an every-fifth-word deletion system controlled for high and low frequency patterns of oral language. Ruddell stated that split-half reliability went as high as .97.[59] He also found correlations of .61 to .78

---

[55] John Bormuth, "Factor Validity of Cloze Tests as Measures of Reading Comprehension Ability," Reading Research Quarterly, IV (Spring, 1969), 364.

[56] Ibid.

[57] Jenkinson, loc. cit.

[58] L.E. Hafner, "Relationships of Various Measures to the Cloze," Thirteenth Yearbook of the National Reading Conference (Milwaukee, Wisconsin: National Reading Conference, 1964), pp. 135-45.

[59] Robert Ruddell, "The Effect of Oral and Written Patterns of Language Structure on Reading Comprehension," Reading Teacher, XVIII (January, 1965), 272.

between cloze scores and the Stanford Achievement Test.

Greene found lower coefficients between cloze scores and total comprehension. A correlation of .51 was found between cloze test scores and the Diagnostic Reading Survey.[60] Weintraub also found high correlations between cloze readability scores and scores on comprehension tests.[61]

Gallant found the reliability of cloze tests to be high, ranging from .90 to .97. These were significant beyond the .01 level of confidence, and her conclusion was that cloze tests were reliable for all three grade levels tested.[62] Bormuth also computed reliability coefficients by two methods and obtained coefficients from .70 to .95.[63]

Kirby designed a study to determine whether cloze test scores would differ significantly from standardized oral and silent reading test scores. The tests used were the Gilmore Oral Reading Test, the Gray Oral Reading Test, and the Gates Oral Reading Test. Statistical differences among these tests were significant for grades one through three. However, for grades five and six, there were no significant

---

[60]Frank Greene, "Modification of the Cloze Procedure and Changes in Reading Test Performances," Journal of Educational Measurement, II (1965), 213-17.

[61]Weintraub, op. cit., p. 2.

[62]Gallant, op. cit., p. 286.

[63]John Bormuth, "Cloze Tests as Measures of Readability and Comprehension Ability" (unpublished doctoral dissertation, Indiana University, Bloomington, 1962), pp. 72-4.

differences among the mean scores. Mean scores on the cloze tests and mean scores on the Gilmore test did not differ at these grade levels. Mean scores on the Gates Oral Reading Test did not differ significantly except for grades one and two. Mean cloze scores did differ significantly from mean test scores on the Gray Oral Reading Test for the entire sample.[64]

In much of the literature on this subject, investigators have deviated from agreed-upon cloze test construction procedures such as deletion rate, response scoring, and other factors of test construction. Because of this, their conclusions are less useful than the conclusions and findings of the above-mentioned studies.

The preceding discussion dealt with the validity and reliability of the cloze procedure. The opinions of many investigators are summed up by Bormuth's statement that "... cloze tests are valuable ... because they are highly reliable and valid, and can be easily and objectively constructed and scored."[65]

Interpretation of Cloze Test Scores

The problem of how to interpret cloze test scores has plagued researchers more than any other facet of the cloze procedure. Initially, raw scores were used, but this method was of little value. Since individual grade levels of a cloze test normally differ in length, raw scores do not lend themselves to comparison. In an effort to eliminate this problem many researchers have converted raw

---

[64] Clara L. Kirby, "Using the Cloze Procedure as a Testing Technique," Educational Resources Information Center Document ED 019 202 (Bethesda, Maryland: ERIC Document Reproduction Service, 1968), Abstract.

[65] John Bormuth, "Comparable Cloze and Multiple-Choice Comprehension Test Scores," op. cit., p. 291.

scores to percentage scores, thus enabling comparisons to be made. Although slightly more useful, percentage scores simply furnish an indication of rank. For example, if a pupil scored 55 per cent on a specific grade level passage, and his classmate scored 45 per cent on the same passage, it is at once obvious which pupil has performed less satisfactorily on the test. Similarly, should a pupil score 60 per cent on a cloze passage of fifth grade difficulty, and make lower scores on each succeeding grade level, it is a safe assumption that the fifth grade passage elicited his best performance. What is not known is how precisely the reader has comprehended the material on any selection. The problem is that conventional measures of reading ability are usually expressed as grade equivalent scores. A method of converting cloze test scores into grade equivalent scores is needed as well as a criterion score to indicate an acceptable performance level on a cloze test.

A logical approach to this problem may be to adopt a criterion score equivalent to a performance level which has traditionally met with acceptance. A survey of the literature reveals a basis for the selection of such a performance level. Authorities in reading have been consistent in accepting the 75 per cent level as an acceptable performance level in reading comprehension. The assumption is made that material on which a pupil is capable of comprehending 75 percent of what he has read is of suitable difficulty to be used for purposes of guided instruction.

Perhaps the earliest mention of the 75 per cent performance criterion was made by Thorndike in 1917. He suggested the use of the 75 per cent performance level and maintained that if a pupil was capable of answering 75 per cent of the comprehension questions covering the reading material, such material was suitable

for use in supervised instructional situations.[66]

Coleman's study attempted to determine which level of paragraph difficulty resulted in the greatest amount of information gain. He used a transparency which contained selected reading passages. While the subjects watched the projected image on a screen, Coleman peeled back the strips of tape which covered the sentences. As the tape was removed in a left-to-right motion, Coleman asked his subjects to guess the succeeding words. The tape was replaced at the completion of this test and the procedure was repeated. Coleman defined information gain as the difference between the subjects' scores on both tests. Maximum information gain occurred on passages on which the subjects scored close to 44 percent.[67] A slight caution is advisable here in that visual memory is likely to have affected these scores.

Among reading authoirities, Betts adopted the 75 per cent comprehension level and it became part of his IRI.[68] Harris, in 1962, suggested the use of this standard;[69] in 1967, Bond and Tinker also supported its use.[70]

---

[66]E. L. Thorndike, "Reading and Reasoning: A Study in Mistakes in Paragraph Reading," Journal of Educational Psychology, VIII (1917), 323-32.

[67]E. B. Coleman and G. R. Miller, "A Measure of Information Gained During Prose Learning," Reading Research Quarterly, III (Spring, 1968), 369-86.

[68]E. A. Betts, Foundations of Reading Instruction (New York: American Book Company, 1946).

[69]A. J. Harris, Effective Teaching of Reading (New York: David MacKay, 1962).

[70]G. L. Bond and M. A. Tinker, Reading Difficulties: Their Diagnosis and Correction (New York: Appleton-Century-Croft, 1967).

Bormuth questioned the adoption of the 75 per cent criterion and stated, after a review of the literature, that:

> The problem is that neither Thorndike nor anyone known to the author has given either logical or empirical reasons as to why these standards should be accepted.
>
> This problem is potentially too serious to be left unattended.[71]

This raises an interesting point of contention. Bormuth obviously felt it was necessary to have empirical evidence to justify a concept, or practice, which has found acceptance by reading authorities and which does, indeed, seem to "work" for teachers and diagnosticians. Furthermore, it seems that this performance criterion has been validated by use. Not content with non-empirical validation, Bormuth designed a series of studies in an attempt to arrive at criterion levels for interpreting cloze tests scores. Interestingly enough, several of Bormuth's studies yielded information similar to Coleman's study. As previously stated, Coleman's study used a modified cloze procedure, and he suggested that maximum information gain occurred at the 44 per cent score. Bormuth's later studies yielded similar information.

In his initial study, Bormuth aimed at the problem of justifying the 75 per cent and 90 per cent comprehension scores (90 per cent comprehension is indicative of the independent level, 75 per cent is indicative of the instructional level) and establishing a frame of reference from which cloze test scores could be interpreted. Bormuth found that pairs of pupils who made scores of less that 17 per cent on cloze

---

[71] John Bormuth, "Empirical Determination of the Instructional Reading Level," Proceedings of the International Reading Association, XIII, Part I (Newark, Delaware: International Reading Association, 1968), 716.

tests exhibited little information gain. Those pairs of pupils whose scores were in the range of 17 per cent to 37 per cent showed a sharp increase in information gain.[72] Pupils whose cloze scores were higher tended to level off with but a slight increase in information gain. One of Bormuth's hypotheses was that the level of maximum information gain on a cloze test, that is, about 37 per cent, is comparable to the 75 per cent level of performance which has been traditionally accepted as the instructional level when ascertained by multiple-choice questions. Similarly, he found that a cloze score of 57 per cent is comparable to the 90 per cent level of comprehension mentioned earlier. Bormuth cautioned against accepting these results based solely upon one study.

Bormuth supported the results he found in his first study with a similar second study. He constructed two reading tests, a cloze and a multiple-choice test, over an identical reading passage. Results indicated that a multiple-choice test score of 75 per cent is comparable to a cloze test score of 44 per cent over the same material. Further, a multiple-choice test score of 90 per cent is comparable to a cloze test score of 57 per cent.[73]

In the next study in this series, Bormuth contended:

> When a student's score falls between 44 and 57 per cent, on one of these tests, the materials are at the level of difficulty thought to be suitable for use in his supervised instruction. Materials on which a student's score is above 57 per cent are suitable for use in his independent study.[74]

---

[72]Ibid., pp. 720-21.

[73]John Bormuth, "Comparable Cloze and Multiple-Choice Test Scores," op. cit., pp. 291-99.

[74]John Bormuth, "Cloze Readability Procedure," op. cit., p. 434.

Culhane and Rankin, in a similar study, reported a cloze test score of 41 per cent to be comparable to the 75 per cent criterion, and a cloze test score of 61 per cent to be comparable to the 90 per cent criterion. Culhane and Rankin concluded:

> In view of these findings, it is now possible for teachers to interpret cloze test scores with some degree of confidence by using specific percentage scores as criteria of acceptable performance.[75]

The scores attained by Culhane and Rankin differed from Bormuth's by four score points. Bormuth's scores of 44 and 57 were actually scores which had been statistically corrected for guessing. This correction was based upon the assumption that an element of guessing enters into a multiple-choice testing situation. Bormuth's conclusion was that uncorrected cloze test scores of 38 per cent and 50 per cent are comparable to the 75 per cent and 90 per cent criteria respectively.

In a fourth study, Bormuth designed a testing situation which is quite similar to the tasks performed by subjects during the administration of an IRI. Using different subjects and materials, Bormuth asked his subjects to read the passages and respond to questions orally. Again, a cloze score of 44 per cent was found to be comparable to the 75 per cent criterion score.[76]

Rankin designed a study to investigate the validity of Bormuth's criterion scores. He concluded that the scores were indeed valid and submitted:

---

[75]J. F. Culhane and E. F. Rankin, "Comparable Cloze and Multiple-Choice Comprehension Test Scores," Journal of Reading, XIII (December, 1969), 193-98.

[76]Bormuth, "The Implications and Use of the Cloze Procedure in the Evaluation of Instructional Programs," op. cit., p. 23.

> Using the 44 per cent criteria (i.e., 75 per cent multiple-choice comprehension), it is clear that there is a fairly close correspondence between results of the formulae employed to check the criteria.[77]

He further stated:

> In using the 58 per cent criterion (rounded up due to test construction concerning fractions of raw scores), we can evaluate the Independent Reading Level by using the cloze procedure.[78]

In summary, it appears that sufficient justification exists in the literature for assuming that a cloze test score of 44 per cent is comparable to a multiple-choice or sentence-completion test score of 75 per cent. And it appears that a cloze test score of 57 per cent is comparable to a multiple-choice or sentence completion test score of 90 per cent. The assumption may safely be made that the kind of reading skills needed during the completion of a cloze test is little different from the kind needed on other types of reading tests. Bormuth's contention on this point is as follows:

> The correlations between cloze readability and conventional comprehension test scores are high, and none of the research has presented convincing evidence that the processes employed in responding to cloze readability tests are, in any major sense, distinguishable from those employed in responding to conventional comprehension tests.[79]

The purpose of many reading tests, notably IRIs and standardized tests, is to give an indication of the instructional level of the pupil. A number of studies have been designed to investigate whether cloze tests are usable for purposes of determining

---

[77] Rankin, op. cit., p. 6.

[78] Ibid.

[79] Bormuth, "The Implications and Use of the Cloze Procedure in the Evaluation of Instructional Programs," op. cit., p. 25.

a pupil's instructional level. A review of these studies follows.

The Use of Cloze Tests to Determine Instructional Level

The initial study in which an attempt was made to determine reading levels by cloze testing was conducted by Ransom in 1965.[80] Ransom administered an informal reading inventory and a cloze test to her population sample and investigated the relationship between the scores. With no research on criterion scores available, Ransom set performance intervals after an examination of pupils' scores. No further justification or rationale was given to validate these intervals empirically. Cloze test percentage scores of 50 and above were considered as being comparable to the independent level of the IRI. Percentage scores which fell in the interval of 30 per cent to 50 per cent were considered as being comparable to the instructional level, and cloze scores below 20 per cent were considered as being comparable to the frustration level.[81]

After applying these performance criteria to the scores, Ransom reported statistically significant correlations between the cloze test scores and the IRI test scores at the instructional and frustration reading levels for all grade levels except first grade. For the first grade level, correlations failed to achieve significance for all three reading levels.[82] Correlation coefficients for the independent reading level were not significant at the .01 level for a majority of the grade levels tested.

It is interesting to note that Bormuth's empirically validated criterion

---

[80] Ransom, loc. cit.

[81] Ibid., p. 39.

[82] Ibid., p. 61.

range of 38 per cent to 50 per cent (not corrected for guessing) is quite similar to Ransom's range which, it appears, was the result of an arbitrary decision.

Kirby's study, which used the same population which Ransom had tested, contained a comparison of scores made by pupils on several standardized oral and silent reading tests with cloze test scores. The performance intervals used by Ransom in her study were also used by Kirby in the analysis of cloze test performance. Kirby stated in her conclusions that "the findings of this study indicate that the use of a cloze test by classroom teachers for determining instructional reading levels of children is a promising technique."[83]

In a study similar to Ransom's, Kirchhoff investigated the relationships between scores on a cloze test and scores on an IRI using a population of first grade children being taught by different methods of reading instruction. Kirchhoff found significant correlations between the cloze test scores and scores on the IRI for all first grade children. In combination, Ransom's and Kirchhoff's studies found significant correlations for all grade levels of the elementary school. It appears that Kirchhoff also used Ransom's criteria for interpreting cloze test scores. Kirchhoff submitted:

> The correlations between the cloze test scores and the informal reading inventory scores at the instructional and frustration reading levels were statistically significant. This close relationship would tend to indicate that a cloze test could aid the first grade teacher in determining the appropriate reading level for children nearing the end of their first grade experience.[84]

---

[83]Clara L. Kirby, "A Comparison of Scores Obtained on Standardized Oral and Silent Reading Tests and a Cloze Test" (unpublished doctoral dissertation, Ball State University, 1967), p. 82.

[84]Leo Kirchhoff, "A Study Utilizing the Cloze Test Procedure to Determine Reading Levels of First Grade Children Who Have Been Taught Beginning Reading by Four Different Approaches" (unpublished doctoral dissertation, University of Kansas, 1968).

Based upon findings of research thus far attempted, the future looks promising for the cloze procedure as a method for determing instructional level.

Summary of Cloze Procedure

The use of the cloze procedure as a method of determining relative readability was first introduced by Wilson Taylor, who contended that the cloze procedure was a superior method of determining readability since it measured a facet of readability called concept load. Taylor's research supported his contention that other readability methods, namely the various readability formulae, could be fooled by literature containing a low vocabulary load with a high concept load. Authorities have accepted cloze tests as valid and reliable instruments capable of measuring general comprehension skills.

Agreement now exists as to the accepted methodology of cloze test construction. A deletion system of every fifth word with standardized blank length has been accepted, although recent studies have submitted that blank length is not crucial to the construction of a cloze test. Reliability coefficients are highest when cloze passages consist of at least 250 words. Scoring cloze tests by the exact-word-response is the most valid method and eliminates any subjective judgment on the part of the examiner. Raw cloze scores are converted to percentage scores to facilitate comparisons among passages of different lengths. Research studies have demonstrated that a cloze score of 38 per cent is comparable to a score of 75 per cent as measured by traditional types of comprehension tasks. If a comparison with multiple choice tests is desired, the score, corrected for guessing, approaches 44 per cent. These scores are indicative of the pupils' instructional levels; scores of 50 and 57 per cent (corrected for guessing) are comparable to the independent level of 90 per cent as measured by traditional types of comprehension tasks.

Three studies indicated the suitability of the cloze procedure as a method

of determining the instructional levels of elementary school pupils.[85, 86, 87]

In reviewing the literature, many research weaknesses become apparent. Many researchers failed to mention the procedures which were employed in cloze test construction. Deletion systems and scoring procedures are often described inadequately. The conceptual difficulty of the cloze materials is rarely mentioned, nor are the readability levels of materials used or the sources of the materials. In studies where cloze tests are correlated with multiple-choice questions, the nature of the questions, their validity, and their development are not described.

Serious sampling errors exist. How subjects are selected is rarely revealed, nor are the actual reading abilities of these subjects given. This latter point is absolutely essential to the interpretation of performance on a cloze test.

More needs to be known about the construct validity of cloze test. Several affective components operating during the completion of a cloze test also need to be investigated. The effects of author's style, passage content which is not congruent with the subject's personal beliefs, and dialectal and syntactical deviations from the subjects' speech patterns should be investigated.

The cloze procedure has contributed a great deal to the areas of both reading and language. It is a promising technique and, with the advent of future research, it can prove to be an invaluable too.

---

[85] Ransom, loc. cit.

[86] Kirby, loc. cit.

[87] Kirchhoff, loc. cit.